BBC RD 1974/28

BBC

RESEARCH DEPARTMENT

REPORT

ACOUSTIC SCALING: subjective appraisal and guides to acoustic quality

A.N. Burd, B.Sc., F.Inst.P., C.Eng., M.I.E.E.
W.N. Sproson, M.A., F.Inst.P.

Research Department, Engineering Division
THE BRITISH BROADCASTING CORPORATION August 1974

BBC RD 1974/28
UDC 534.846.6

ACOUSTIC SCALING: subjective appraisal and guides to acoustic quality

A.N. Burd, B.Sc., F.Inst.P., C.Eng., M.I.E.E.
W.N. Sproson, M.A., F.Inst.P.

Summary

The assessment of changes made in an acoustic model requires a group of subjects to listen to music or other programme material recorded at each stage of the model work and to compare the qualities of the sound obtained. Difficulties have been found in obtaining consistent judgements and various techniques of subjective appraisal have been tried. The preparation and presentation of material has been lengthy and the analysis of the results even more so, but it is possible that, from the data obtained, some indications may be found of the important factors which together make up 'acoustic quality'.

Issued under the authority of

P. Lamaq

Head of Research Department

Research Department, Engineering Division,
BRITISH BROADCASTING CORPORATION

(PH-119)

Foreword

This is one of a series of reports on acoustic modelling and deals with the same matter as Reference 2. Whereas Reference 2 deals in some detail with the precise changes made, the reason for making them and the results obtained, the present report concerned primarily with techniques of appraisal including subjective testing techniques and the mathematical analysis of the results

ACOUSTIC SCALING: SUBJECTIVE APPRAISAL AND GUIDES TO ACOUSTIC QUALITY

ACOUSTIC SCALING: SUBJECTIVE APPRAISAL AND GUIDES TO ACOUSTIC QUALITY

A.N. Burd, B.Sc., F.Inst.P., C.Eng., M.I.E.E.
W.N. Sproson, M.A., F.Inst.P.

1. Introduction

The work on acoustic modelling, carried out in the BBC Research Department during recent years, has had several distinct aims, of gradually widening application. In the first part of the work[1] the model of an existing orchestral studio was shown to be capable of reproducing the acoustic quality of the real studio as assessed from the programme signals obtained. It was stated by a group of experienced listeners that although the two sounds were not identical, sufficient similarity existed to permit valid judgements to be made. In the second phase of the work[2] a series of possible changes to the studio (reflectors, diffusers, absorbers in various dispositions, etc.) were modelled and the resulting changes in the acoustic quality evaluated. Subsequent uses of the model have tended progressively towards achieving increased understanding of acoustic quality and less towards immediate practical applications. Thus, the roof height has been increased to an extent which must be economically out of the question (in a real studio) and the model has been applied to a study of factors that are considered to be important in the determination of acoustic quality.

Throughout all these studies the question of subjective appraisal has been of paramount importance. It was thought, in the early days of modelling, that by permitting a subjective evaluation of the overall acoustic quality it was possible to avoid the unresolved question as to whether, in objective terms, the quality in one case was better than in another. In the first stage proving experiment the decision that two recordings were similar was comparatively straightforward. Judgements since then have required a 'better than' or 'worse than' decision, which has disproved the above simplistic view of modelling assessment. However, it is considered that, in the long term, the work described in this report will contribute to a greater insight into the judgement of acoustic quality which, in turn, will lead to the development of objective measurements that permit the prediction of subjective assessments.

2. Choice of observers

Previous exercises involving subjective assessment have often concentrated on the opinions of professional engineers because such subjects are always available within a broadcasting organisation. Where the preferences of the listening public are being examined, it is obviously necessary that the subjects be representative of them. Thus the experiments on 'Listener's Sound-Level Preferences'[4] compared the views of the ordinary listeners with those of engineers and of musicians and found different preferences for the three groups.

During a study of preferred microphones balances carried out in 1966,[5] the consistencies of judgement of several classes of subject were compared. It was found that research engineers and sound balance engineers, while capable of assessing critically the technical quality of programmes, did not agree on the ranking of a set of different microphone balances. Concert-going members of the public, on the other hand, proved to be highly consistent and agreed in preferring one particular type of balance.

Some previous work on acoustic scaling[6] used blind subjects, whose critical faculties are perhaps more fully developed with regard to certain aspects of sound quality in the absence of a view of the area involved. While this may be true for judgements of size, shape or proximity to a surface, it seems unlikely 'a priori' that it extends to musical appreciation.

Since it was apparent that subjective assessments were going to be required at several stages of the acoustic scaling study, it was decided to use only broadcasting balance engineers having a particular interest in musical work. As has been shown previously, such subjects are capable of critically assessing technical quality although they may not agree amongst themselves on preferred quality.

3. Assessment of changes to an acoustic model

3.1. Simple overall assessment

The modelling study had its first application in the examination of a number of possible modifications that could be applied to the orchestral studio.[2] In each of these conditions a recording was made of an excerpt of anechoic music reproduced over two loudspeakers and picked up by a pair of spaced omnidirectional microphones. A preliminary selection of eight preferred conditions was made by those carrying out the study and these recordings were presented to several groups of listeners. This choice may have influenced the results obtained.

The subjects were asked to listen to a side-by-side presentation of the signals provided by a modified condition compared with those obtained using the standard unmodified condition. The two recordings were run in approximate synchronism and a key was operated to change from one to another. A visual indication was given of the programme ('A' or 'B') being heard at a particular instant.

The subjects were asked to rate each of the changed conditions relative to the standard condition with a rating score of +5 to −5 depending on whether there was an overall improvement or impairment; there was a space on the questionnaire for comments.

The results obtained from eight observers were averaged to obtain an overall judgement. For six different modifications the average value was not significantly dif-

ferent from zero; however, an examination of the results and the comments made it clear that this figure had not occurred because the observers thought all the modifications were alike. Considerable differences were noted which by some were rated an improvement, while others considered the same condition to be a degradation. The comments showed that while one person's judgement might be affected considerably by, say, an improved string tone, another would be more concerned over a simultaneous change in the low-frequency reverberation.

If a single arbiter of overall quality existed, then a simple assessment might be a possibility. Since, however, many views have to be considered, a closer examination of the factors that influence them was necessary.

3.2. Presentation of material

As soon as the subjects were asked to make more critical judgements, they become increasingly aware of those parts of the presentation which hindered easy comparison. Various possible methods of presentation were considered in order to select that which most assisted the observers; the success of a particular method could be assessed by the number of correct judgements obtained in a test where the standard condition was compared with itself.

It is doubtful whether comparison material could be presented simultaneously on two headphones; changes of loudness, for instance, could be judged in this way but probably not changes of quality. The simultaneous presentation of two samples over loudspeakers, analogous to the presentation of two colour samples for visual assessment, is clearly pointless.

Thus a test relies on the subject's auditory memory in effecting a comparison of two samples presented at separate times. Since such a memory is probably fairly short, this places an upper limit on the useful length of a sample; on the other hand, it must be long enough to be representative of the programme material. Clearly the pauses between presentations should also be kept to a minimum. Discussions and simple tests placed the optimum length between ten and fifteen seconds and samples of this duration are normally presented.

The simplest possible presentation is a fifteen second excerpt of material A followed by the same excerpt of material B; this allows a single comparison of this excerpt. However, this method may be improved by presenting the material in the form of A-B-A when a judgement based on the first comparison of B with the memory of A can be confirmed – or rejected – on the basis of the subsequent repeat of A. This is the standard method used, heretofore, in the majority of subjective tests.

The simultaneous availability of the two samples of material allows the subject or the experimenter to switch at will between samples A and B. For groups of subjects (groups, necessarily, to carry out the large number of comparisons for many qualities with the expenditure of a reasonable amount of experimenter's time), this meant that one person defined the switching points in a way which could never suit all those involved. Additionally it was always frustrating to find that a change in the scoring of the music or other musical factors invariably occurred at the chosen instant of time and rendered a comparison more difficult (see Section 4.2).

A form of presentation has therefore been evolved in which a short excerpt (say ten seconds' duration) of A is followed by the same excerpt of B, and then a further repetition of A and B; a different passage of music then follows, the same A-B-A-B pattern being retained. This may be repeated as often as is thought necessary or for the greatest number of passages possible from the material available.

The first application of this form of test proved as unsatisfactory as previous forms because of insensitive divisions of the material into short samples; to musically sensitive people incomplete musical phrases and abrupt starts and stops were distracting. However, when the assistance of a professional recording and balance engineer was enlisted, a test was prepared which met most requirements. Musically complete phrases of duration 10 – 20 seconds were faded in and out, the phrase being chosen so that the end flowed smoothly back into the beginning without changes of key, tempo, etc. Announcements of sample A or B were 'voiced over' the end/start of each sample and the whole was considered most satisfactory (see Section 4.3).

3.3. Method of scoring

As described in Reference 7, subjects can be asked to indicate by a mark on a line the particular value they associate with a given quality having two extremes (±5 grades) defined by words. On the basis of preliminary experiments ten independent qualities were selected which appeared to be the most significant. The form of questionnaire is shown in Appendix 1. Thus a completed questionnaire for one comparison test will have eleven marks on it and each mark is translated into a quantitative rating ranging from −5 through 0 to +5 by simple linear scaling. The 0 central point implies that the unknown B excerpt is regarded as identical with the original A version for the particular quality under consideration. The questionnaire includes an overall assessment on a like/dislike basis and this is assumed to be largely dependent on the ten individual qualities (tonal warmth, definition or clarity, colouration etc.) used in the questionnaire.

The amount of information obtained by scoring a large number of qualities is considerable, but the subjective tests are commensurately more difficult. It was felt that the potential gains were worth pursuing provided the degree of difficulty did not prejudice the validity of the results.

Each condition was compared with the original, using recordings made in the model, and one presentation was of the original recording compared with itself. Subjects were not informed that this was being done, and, although experienced subjects would expect such a test to be included, they had no way of predicting which of the tests it

would be. Such a check permitted a selection of consistent subjects to be made. A subject was judged to be consistent if he scored less than ±0·5 for 10 or the 11 qualities of the comparison of the original recording with itself and he must also be able to note some real differences as greater than ±0·5; if this latter proviso were not included, then a subject who never scored more than ±0·5 at any time would be considered to be consistent.

3.4. Factor analysis (multi-dimensional scaling)*

The use of factor analysis in examining 'subjective acoustic experience in concert auditoria'[7] has been described by Hawkes and Douglas. Briefly, factor analysis attempts to discover the minimum number of independent (mathematically orthogonal) components – called factors – that are required to describe adequately the original data. Thus, if N sets of results using n original independent variables are obtained in a series of subjective studies, they may be described in an n dimensional space (by regression analysis, for example). Factor analysis seeks to reduce the number of variables (n) by formulating new mathematically orthogonal variables (i.e. factors) which describe the original data in a smaller number of variables m where $m \leqslant n < N$. Factor analysis (principal components method) operates by determining the latent roots of an n x n matrix of correlation coefficients and these are used in descending order of magnitude. The latent roots are also the variances of the new variables (i.e. factors) and for n variables the sum of the latent roots is also n. It is therefore possible to determine the number of factors which will contain, for example, 90% of the total variance. It frequently happens that the first two or three factors contain most of the variance and the whole of the data can thus be substantially described with a small number of factors. Expressed in other words, a plot of variance against number of factors shows a point at which almost all the variance is accounted for and no further factors are needed to explain the original data. It was felt that the approach described by Hawkes and Douglas could probably provide the information that was required and, after informative discussions with Hawkes, the technique was used to examine, separately, two large music studios.[8] It has been felt subsequently that the use of this technique to examine a single studio may not be appropriate but its suitability to examine modifications which affect different qualities to varying extents is not in question.

4. Application of the method

4.1. 'Confusion matrix'

One approach to multi-dimensional scaling has been described by Hawkes.[9] In this method short excerpts of the available material are presented to the subject two samples at a time. The subjects are asked to say whether the second sample of a particular pair is a repetition of the first, or is different. Each particular pairing is repeated many times and, on the basis of the number of replies in which the subject fails to make the correct identification, the experimenter is able to judge the extent to which the two samples are different, i.e. greater similarity leads to greater confusion and thus to a greater number of incorrect responses. The individual samples can now be placed in order of 'difference' and, in addition, the knowledge of the degree of difference between each pair allows a measure of their position in a multi-dimensional space to be made.

Since Hawkes was interested in exploring the application of this technique, he offered to use the recordings produced in the BBC model studies and to test them using his own subjects. Short excerpts from the original recordings were transcribed onto discs so that any pair could be simply selected by the experimenter and presented in sequence with the minimum of delay.

Analysis of the results showed clearly that mathematically four factors were required to explain the subjective assessments. However, it is inevitable with this technique that no indication is given of the nature of the significant factors, which must be determined independently.

An unsuccessful attempt was made to locate the factors subjectively. For each factor three recordings were selected which represented the maximum, minimum and an intermediate value of that factor; these three recordings were listened to in order of increasing amounts of the factor in the belief that its presence should become increasingly obvious. In one case it was thought probable that the factor was the variation in background noise on the recordings, which probably resulted in part from progressive improvements to the equipment during the course of the experiments. Other factors were not clearly identified.

The simplicity of this type of subjective test is attractive but its sensitivity to unintentional variations is a limitation to its usefulness. A minor variation to this approach presents the data in all possible sets of three. The subjects are asked to state in each case which pair is most alike and which pair is most different, a slightly more complex assessment. More information is obtained at one time but the possibility of errors of judgement would appear to be greater. The data obtained is handled in exactly the same way as previously described and the results are (or should be) the same.

4.2. 1971 Subjective tests

For the subjective tests carried out in 1971, seventeen subjects were available. The material consisted of seven comparisons which were obtained from recordings made in the model after each acoustic modification; the changes are described briefly in Table 2 (see below) and in more detail in Reference 2. The material was recorded on two separate magnetic recording tapes and replayed from tape machines run in approximate synchronism. One person was responsible for the switching operation to change from material A to B. The disadvantages of this type of presentation have been discussed above.

* 'multi-dimensional scaling' is a statistical method of analysing complicated data and bears no relation to 'acoustic scaling' which forms the main topic of this report.

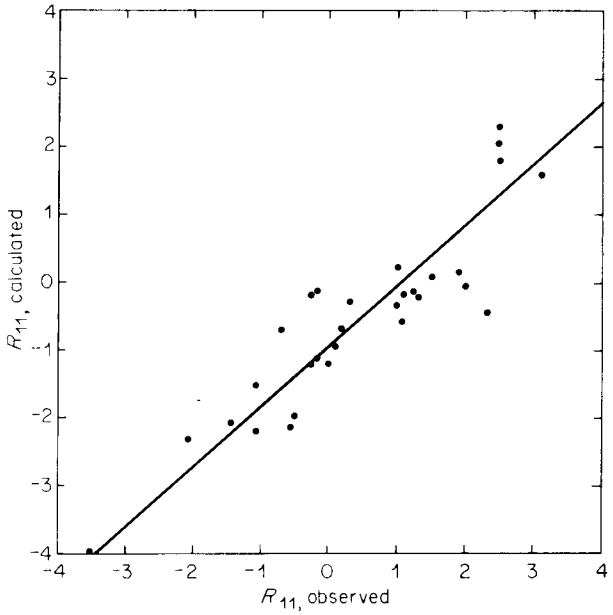


Fig. 1 - 1971 Tests: relationship between overall assessment (R_{11}) calculated from regression analysis and the directly observed value

Applying the criterion mentioned above, only two of the subjects were consistent. The results of a further two subjects who were jointly responsible for the quality of much of the output from BBC Radio were also examined in some detail, notwithstanding the fact that they had been rated inconsistent.

The results from these four subjects were subjected to linear regression analysis and the following 'best fit' equation was produced:—

$$R_{11} = 0{\cdot}3887R_3 + 0{\cdot}0466R_8 - 0{\cdot}3225R_9 + 0{\cdot}4872R_{10} - 0{\cdot}35 \quad (1)$$

where R_{11} = overall assessment
R_3 = colouration
R_8 = timbre
R_9 = brilliance
R_{10} = string tone

The multiple correlation coefficient, *r*, has a value of 0·8988. A plot of the relationship between the calculated R_{11} and the actually observed R_{11} is shown in Figure 1. This shows fairly good agreement between the calculated and actual values of R_{11} but there are errors of estimation of up to 1·8 units. The r.m.s. error of estimation is ±0·64 units.

From the above analysis, it would appear that colouration, timbre, brilliance and string tone play an important part in accounting for the overall assessment.

4.3. 1972 Subjective tests

For this series of tests the improved form of presentation described in Section 3.2 was used. 21 subjects were available and it was a measure of the improved presentation that eight of the subjects achieved consistency as defined in Section 3.3.

Regression analysis of the results of the eight consistent subjects gave the following equation.

$$R_{11} = 0{\cdot}4163R_1 - 0{\cdot}0898R_2 - 0{\cdot}1392R_3 + 0{\cdot}1241R_6 - 0{\cdot}6442R_8 + 0{\cdot}5755R_9 + 0{\cdot}0695R_{10} + 0{\cdot}2864 \quad (2)$$

where R_1 = tonal warmth
R_2 = definition
R_3 = colouration
R_6 = intimacy
R_8 = timbre
R_9 = brilliance
R_{10} = string tone
R_{11} = overall assessment

This is a more complicated relationship than that obtained in the 1971 subjective tests. It included the four variables previously included (viz. R_3, R_8, R_9 and R_{10}) but adds R_1, R_2, R_6 to the list. R_1 (tonal warmth) has a relatively large coefficient of 0·4163. The multiple correlation coefficient is 0·8973 which is virtually identical with that obtained in the 1971 subjective tests (0·8988).

An assumption which is made in producing Equations (1) and (2) is that although observers may differ to some extent in their judgements of individual qualities (and also in the overall assessment) the manner in which they form their overall assessment from the individual qualities is substantially independent of the observer. Some justification of this assumption is obtained in an analysis of the results, observer by observer, which shows fairly direct evidence that significant similarities exist between individual observers. Further, if there were not some important elements held in common by different experienced listeners, judgement of acoustic quality would become a very individual as well as a subjective matter and all attempts at a statistical analysis of any kind would be doomed to failure. There is, however, considerable evidence of broad agreement between skilled (consistent) observers and this acts as an encouragement to pursue the available data analytically.

Factor analysis of the results using a computer programme based on the method of principle components[9] showed clearly the existence of four factors which accounted for most of the variance. Additional information given by the analysis defined the weighting for each *factor* (F) of each of the *original* qualities (Q). Thus it was found that Factor 1 contained Qualities 2, 3, 8, 9 and 10 to a significant extent and the others less importantly. The appropriate weightings were:—

$$F_1 = -0{\cdot}863Q_{10} + 0{\cdot}852Q_3 + 0{\cdot}843Q_8 - 0{\cdot}837Q_9 + 0{\cdot}783Q_2 \quad (3)$$

where Q_{10} = String tone)
Q_3 = Colouration) expressed in statistical '*t*' units,
Q_8 = Timbre) i.e. $(R_{10})_t$, $(R_3)_t$ etc., using the
Q_9 = Brilliance) notation of Equation (7) below.
Q_2 = Definition

This factor correlated strongly with the overall assessment (r = 0·84), and was the only factor which did so.

Factor 2 contained predominantly Q_1 and Q_7;

$$F_2 = 0{\cdot}872Q_1 - 0{\cdot}815Q_7 \qquad (4)$$

where Q_1 = Tonal warmth)
Q_7 = Hardness) expressed in statistical 't' units

Factor 3 contained predominantly Q_5 and Q_6;

$$F_3 = 0{\cdot}915Q_5 + 0{\cdot}695Q_6 \qquad (5)$$

where Q_5 = Liveness)
Q_6 = Intimacy) expressed in statistical 't' units

Factor 4 contained predominantly Q_4;

$$F_4 = 0{\cdot}754Q_4 \qquad (6)$$

where Q_4 = fullness of tone in 't' units.

It will be noted from the above table that the important qualities do not appear more than once in the four factors.

The same computer programme was also run with all the available data from the twenty-one observers, but the results did not show the same relatively simple relationships as described above and no deductions were made from this latter analysis.

The results of regression analysis have already been indicated to a limited extent in Section 4.3 but it is convenient to compare the results for the 1971 and 1972 tests in Table 1.*

This shows that the four variables in the 1971 tests are repeated in 1972 but the magnitudes of the coefficient have changed and quality 1 (tonal warmth) is added to the variables with a relatively large coefficient. R_{10} (string tone) appears to assume much less importance in the 1972 tests than was indicated in 1971. The negative coefficient for R_9 (brilliance) is taken to imply that Maida Vale 1, in its reference condition, has too much of this quality and some reduction is desirable. It is to be noted, however, that this disagrees with the results shown in Table 1 of Reference 8.

* Some differences in sign will be observed between coefficients in Table 1 and the corresponding terms in Equation (2). A different convention was used in 1972 but to make the comparison easy, Table 1 uses the 1971 convention. Polarities are explicitly stated in Table 2.

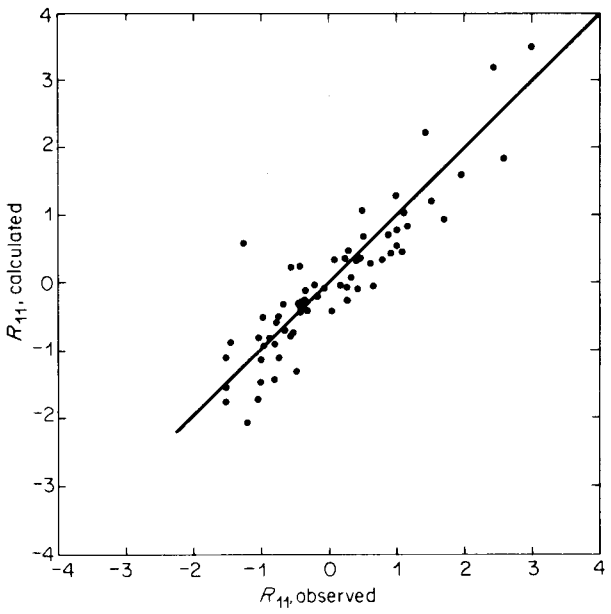


Fig. 2 - 1972 Tests: relationship between overall assessment (R_{11}) calculated from factor analysis and the directly observed value

The equation of best fit derived from factor analysis is

$$(R_{11})_t = 0{\cdot}243F_1 - 0{\cdot}097F_2 - 0{\cdot}337F_4 \qquad (7)$$

Where F_1, F_2 and F_4 are factors defined previously and the $(R_{11})_t$ implies the overall assessment expressed in statistical 't' units (zero mean value and unit standard deviation). The multiple correlation coefficient given by this analysis is 0·9002 which is virtually the same as that given by regression analysis: F_1 accounts for most of the correlation as commented in Section 4.3. Fig. 2 shows $(R_{11})_t$ as calculated from Equation (7) plotted against the directly observed data for R_{11} translated back into the original scalings so that it is directly comparable with Fig. 1. Although Equation (7) and Fig. 2 show a hopeful situation, a real difficulty still exists in terms of naming and understanding what the factors F_1, F_2 and F_4 mean in physical or psychophysical terms. F_1, for instance, combines a number of diverse qualities (string tone, colouration, timbre, brilliance and definition), which are not obviously embraced by any one single well-known acoustic quality or concept. A number of analyses of other subjective assessments have resulted in F_1 having high correlation with over-

TABLE 1

Coefficient in Regression Equations

	R_1	R_2	R_3	R_6	R_8	R_9	R_{10}	Multiple Correlation Coefficient
1971			0·3887		0·0466	−0·3225	0·4872	0·8988
1972	0·4163	+0·0898	+0·1392	0·1241	+0·6442	−0·5755	0·0695	0·8973

TABLE 2
Assessments of Modifications to Model of Maida Vale Studio 1
Mean Values
Modification and Brief Description

Quality	Date of test	1 empty studio	2 patches of absorbent on end wall	3 absorbent on roof	4 curved canopy	5 choir seating removed: orchestra to end	6 plastic diffusers in roof	7 hard ceiling	8 reference condition (check)
1. warmth	1971	−0·20	0·20	0·375	0·325	0·45	0·225	0·575	0·0
	1972	0·55**	0·275	0·562*	0·312	0·325	0·175	0·375	−0·05
2. definition	1971	−0·50	−1·125**	−1·575*	−1·175	−0·25	0·6	−0·375	−0·35
	1972	−1·138**	−0·475	−1·862**	−0·238	−0·388	0·138	−1·312**	0·038
3. colouration	1971	−0·40**	0·60	−1·5*	0·025	−0·1	−0·575	−0·225	−0·175
	1972	−0·70*	−0·412	−2·225**	−0·050	0·212	0·212	−0·825	−0·212
4. fullness of tone	1971	0·10	0·225	0·525	0·0	0·425	0·275	0·8*	−0·175
	1972	0·512	0·425	0·825*	0·425	0·962*	0·350	0·175	−0·062
5. liveness	1971	0·725	0·0	−0·225	−0·225	0·775	0·75	−0·35	−0·2
	1972	0·262	0·212	−0·438	−0·150	0·70	−0·062	0·325	0·0
6. intimacy	1971	−0·07	0·0	−0·75	0·175	0·2	0·8	−0·5	0·15
	1972	−0·45	−0·138	0·038	0·675*	+0·488	−0·050	−0·512	0·025
7. hardness	1971	0·75	0·85**	2·075*	1·05	−0·075	−0·125	−0·10	−0·125
	1972	0·525	0·75	1·138	0·388*	−0·100	−0·025	0·338	−0·138
8. timbre	1971	−0·325	−0·275	−1·5	−0·35	0·1	0·2	0·15	0·075
	1972	−0·038	0·050	−1·312**	−0·038	−0·25	0·338	−0·412	0·0
9. brilliance	1971	0·475	0·80	1·775	0·9	0·025	−0·15	1·075	−0·275
	1972	0·500**	0·388	1·4**	−0·15	0·1	−0·088	1·212	−0·150
10. string tone	1971	0·675	−0·275	−1·575	−0·05	0·65	−0·5	−0·625	−0·05
	1972	0·388	−0·275	−0·775*	0·125	−0·062	0·275	−0·375	0·025
11. overall	1971	−0·8*	−0·375	−2·35*	0·375	0·2	−0·525	−1·1	−0·05
	1972	−0·7**	−0·31	−2·29**	−0·38	−0·45	−0·06	−1·18	0·01

* result significant at 5% level
** result significant at 1% level

Polarities	Quality 1	2	3	4	5	6	7	8	9	10	11
Negative	colder	muddier	more coloured	thin	dead	less	hard	poorer	more brilliant	worse	dislike
Positive	warmer	clearer	less coloured	full	live	more	mellow	better	duller	better	like

all assessment and it would appear that, to a first approximation, F_1 indeed is, or is very closely associated with, overall assessment. If this is true, it unfortunately does not assist in better understanding or indicate methods of devising objective techniques of measurement which could ultimately replace skilled observers. The factors F_2 F_3 and F_4 are easier to comprehend as the number of qualities are less and they are more obviously related to one another.

5. Discussion of results

The results that have been quoted above are only used to exemplify the type of analysis that can be employed and the form of the results obtained. A further statement of the experiments involved and the conclusions reached are contained in another report.[2]

However, it is interesting to compare the results obtained by the consistent observers in the two series of tests. Table 2 shows the mean values obtained for all the individual qualities for both 1971 and 1972 tests, using the 8 consistent observers of the 1972 tests and the four selected observers from the earlier tests. There is a considerable measure of agreement although some contradictions are seen. If a change of less than 0·5 grade is taken as agreement between the two sets of results, then 24 of the total of 88 pairs disagree (i.e. 64 pairs of results are in substantial agreement).* The significance of each result has been assessed by Student's '*t*' test and, of the 1971 results, 9 are significantly different from the reference condition (rated at zero), with 3 at the 1% level and 6 at the 5% level. For the 1972 results, 18 are significant, with 10 at the 1% level and 8 at the 5% level. This increase from 9 to 18 significant results is felt to be a direct consequence of the improved experimental presentation. Five of these significant results are common to both the 1971 and 1972 tests. Of the two pairs of results where the mean results differ by more than 0·5 grade, 14 involve a change of polarity and these are perhaps more disturbing than the others. For example, a slight preference on overall assessment of 0·375 for the curved canopy variant in 1971 has changed to a dislike of about the same magnitude (−0·380) in 1972. It is only fair to point out that in this case neither result is significant in its own right. In fact, in no case of a change of polarity is either of the pairs of results significant, with one exception and in this case only one of the pair is significant. (This exception is quality 1, tonal warmth for the empty studio).

For an assessment of the difficulty and complexity involved in these modelling experiments, it is felt that Table 2 reveals a fairly satisfactory state of affairs although some of the contradictions would cause one to hesitate before asserting conclusions too positively.

It was disappointing, but not altogether unexpected, that the analysis of all the subjects' results (non-consistent as well as consistent) by the method of factor analysis failed to show a simple pattern. The preliminary simple-preference tests showed that, at a conscious level, subjects attached a varying importance to different qualities. Thus even the discovery of a concensus opinion from the consistent observers could not be predicted. Perhaps the work reported here could lead to the establishment of a panel of selected observers who will define the BBC acoustic quality, although this is not the long-term aim of this work.

Further subjective tests are now planned in which new and improved recordings will be evaluated. If the results of these confirm the findings of the 1972 tests, then greater confidence will be placed in the conclusions.

6. Disclaimer

As was mentioned in Section 3, a selection from the 50-odd available recordings was made in order to reduce the tests to manageable proportions. This selection bore in mind the extremely limited sum of money available for any modifications to the acoustic treatment of the full-size studio and conditions in which only a small change was noted were inevitably chosen. This unfortunate limitation has meant that subjects were assessing very small differences in any individual quality.

During the course of the work the direction of the research changed from a simple selection of a suitable modification for a particular studio to a study having much wider implications.

Two facts were responsible for the change of direction of the research. The failure of a simple assessment to provide the answer led to unexpected complication of the work while practical considerations defined the acceptable change in the studio, and a different solution from those envisaged in the early work was carried out. Modelling showed this change to give an improvement in the acoustic quality from the studio and the results of this comparison have been described elsewhere.[2]

Nevertheless, it was realised that the recordings provided a unique opportunity to explore the ways in which an individual subject defines acoustic quality. With hindsight it is obvious that greater changes should have been used to enable more clear-cut decisions to be reached by the subjects. The opportunity will occur again in some current work on the effects of variation of roof height[3] and further tests will be prepared and assessed by subjects.

7. Conclusions

It was originally thought that acoustic modelling would provide a relatively simple way of assessing the overall acoustic quality of a music studio. This does not seem to be the case and more sophisticated technqiues have had to be used. These appear to be yielding useful results and two of the experiments have shown that four factors are required to describe acoustic quality. In one case (confusion matrix, Section 4.1) the variance is completely described by the four factors although identification of the

* Bearing in mind that the last column is a check, it would probably be more correct to say that 53 pairs out of 77 show 'agreement' and 24 'disagree' (as defined in the text).

individual factors has not been possible. In the other case (1972 Subjective tests, Section 4.3) four factors accounted for 82% of the variance and the individual factors are identified. Further work is planned which, hopefully, will confirm the results described in this report. Multi-dimensional scaling certainly appears to be capable of helping the search for factors which contribute to acoustic quality, but it has been shown that sensitive presentation of material is essential if meaningful results are to be obtained. Under musically satisfying conditions, consistent results can be given by a number of observers.

8. References

1. HARWOOD, H.D., BURD, A.N. *and* SPRING, N.F. 1972. Acoustic scaling: an evaluation of the proving experiment. BBC Research Department Report No. 1972/3.

2. BURD, A.N., HARWOOD, H.D., LANSDOWNE, K.F. *and* HUGHES, S.A. Acoustic modelling: examination of possible modifications to Maida Vale Studio No. 1. BBC Research Department Report in course of preparation.

3. HARWOOD, H.D. *and* LANSDOWNE, K.F. 1974. Acoustic scaling: the effect on acoustic quality of increasing the height of a model studio. BBC Research Department Report No. 1974/12.

4. SOMERVILLE, T. *and* BROWNLESS, S.F. 1949. Listener's sound-level preferences. BBC Quart., 1949, **III**, 4, pp. 245 – 250.

5. JONES, D.K. 1966. A subjective investigation into preferred microphone balances. BBC Research Department Report No. B-090, Serial No. 1966/25.

6. KRAUTH, E. 1960. Klanggetreue Nachbildung der Raumakustik durch Modelle. Dr.-Ing. Thessis, Technische Hochschule, Munich, 1960.

7. HAWKES, R.J. *and* DOUGLAS, H. 1970. Subjective acoustic experience in concert auditoria. Acustica, 1971, **24**, 5, pp. 235 – 250.

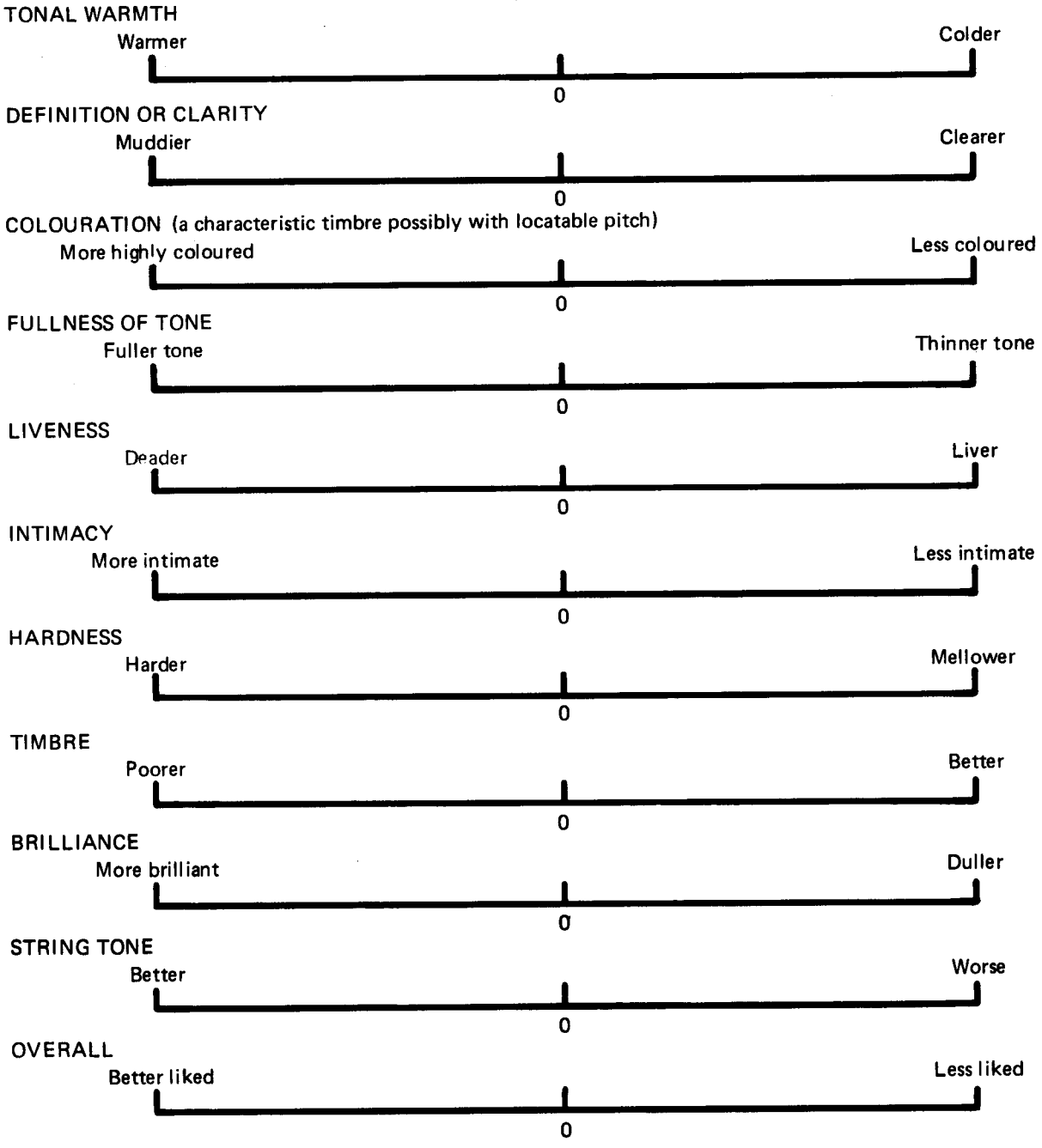
8. SPROSON, W.N. Subjective study of two large music studios. BBC Research Department Report No. 1974/9.

9. HAWKES, R.J. 1970. Multi-dimensional scaling: a method for environmental studies. Building, 1970, **25**, pp. 69 – 72.

10. LAWLEY, D.N. *and* MAXWELL, A.E. 1963. Factor analysis as a statistical method. London, Butterworth, 1963, pp. 45 – 54.

SMW/VK

Appendix 1



Name: .....

Item No. .....

Date .....